Abstract

MapReduce is a programming model used by Google to process large amount of data in a distributed computing environment. It is usually used to perform distributed computing on clusters of computers. Computational processing of data stored on either a file system or a database usually occurs. MapReduce takes the advantage of locality of data, processing data on or near the storage areas, thereby avoiding unnecessary data transmission. The simplicity of the programming model and the automatic handling of node failures hiding the complexity of fault tolerance make MapReduce to be used for both commercial and scientific applications. As MapReduce clusters have become popular these days, their scheduling is one of the important factor which is to be considered. In order to achieve good performance a MapReduce scheduler must avoid unnecessary data transmission. Hence different scheduling algorithms for MapReduce are necessary to provide good performance. This paper provides an overview of four different scheduling algorithms for MapReduce namely; Scheduling algorithm in Hadoop, Longest Approximate Time to End (LATE) MapReduce scheduling algorithm, Self-Adaptive MapReduce (SAMR) scheduling algorithm and Enhanced Self-Adaptive MapReduce scheduling algorithm (ESAMR). An overview of these techniques is provided through this paper. Advantages and disadvantages of these algorithms are identified.
Survey on MapReduce Scheduling Algorithms

References

- Xiaoyu Sun, Chen He and Ying Lu "ESAMR: An Enhanced Self-Adaptive MapReduce Scheduling Algorithm," (2012) IEEE 18th International Conference on Parallel and Distributed Systems.

Index Terms

Computer Science

Algorithms

Keywords
MapReduce; Programming model; Scheduling algorithms;